BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 22 2005, pages 4196–4197
doi:10.1093/bioinformatics/bti675

Data and Text Mining

METIS: multiple extraction techniques for informative sentences

A. L. Mitchell[1,2,*], A. Divoli[1], J.-H. Kim[3], M. Hilario[3], I. Selimas[1] and T. K. Attwood[1,2]

[1]Faculty of Life Sciences and School of Computer Science, University of Manchester, Oxford Road, Manchester M13 9PT, UK, [2]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [3]Artificial Intelligence Laboratory, University of Geneva, CH-1211 Geneva 4, Switzerland

Received on June 29, 2005; revised on August 31, 2005; accepted on September 8, 2005
Advance Access publication September 13, 2005

ABSTRACT

Summary: METIS is a web-based integrated annotation tool. From single query sequences, the PRECIS component allows users to generate structured protein family reports from sets of related Swiss-Prot entries. These reports may then be augmented with pertinent sentences extracted from online biomedical literature via support vector machine and rule-based sentence classification systems.

Availability: http://umber.sbs.man.ac.uk/dbbrowser/metis/

Contact: mitchell@ebi.ac.uk

Supplementary information: http://umber.sbs.man.ac.uk/dbbrowser/metis/supp_inf_results.html

1 INTRODUCTION

There is a pressing need for computational tools to facilitate annotation of sequence data, a task that, for each sequence or set of sequences, involves culling information from various sources, including the literature. A major challenge for such tools is to trace pertinent papers and to extract relevant information from them. Several automated approaches tackle the information extraction problem [e.g. PASTA (Gaizauskas *et al*., 2003) for protein structure and MedMiner (Tanabe *et al*., 1999) for gene expression profiling], but these tools have a specific focus that is not directly applicable for database annotation.

To this end we have developed METIS, building on an existing annotation tool PRECIS (Mitchell *et al*., 2003) that automatically creates protein reports from related entries in Swiss-Prot [the manually annotated component of UniProt (Apweiler *et al*., 2004)]. Although PRECIS gathered the linked literature from each entry, it never directly exploited that information. An innovation in METIS is to use the data in the Swiss-Prot entries to find relevant literature, or to find search terms with which to seek this out. The literature (in the form of abstracts) is then collected and passed to two sentence classification components that extract informative sentences and present them to the user.

2 IMPLEMENTATION

Figure 1 shows an overview of METIS. The software takes as input a FastA format sequence or a Swiss-Prot identifier. The PRECIS component performs a BLAST (Altschul *et al*., 1997) search of Swiss-Prot and digests the related entries to create a structured report that details protein structure, function and disease, keywords, and database and literature cross-references.

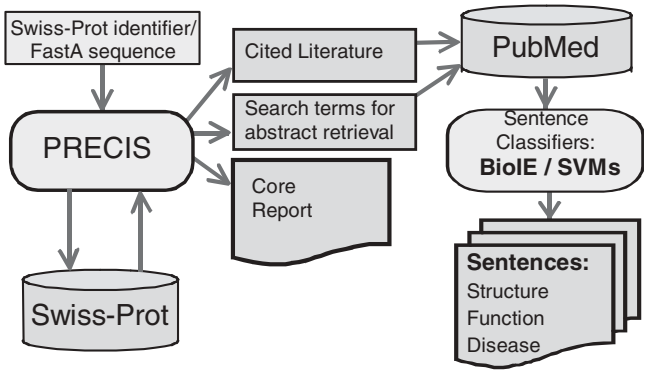

Fig. 1. Flow chart showing the sequence of actions performed by METIS.

Using PubMed identifiers from each Swiss-Prot entry, corresponding abstracts are retrieved and passed to the sentence classifiers. Refineable PubMed query terms are also produced by analysing the Swiss-Prot entries. These allow users to perform wider literature searches and to run the sentence classifiers on the output.

The first sentence classification component is a set of support vector machines (SVMs), built as part of the BioMinT text-mining project. It was developed on three specialized corpora for structure, function and disease, totalling 2406 positive and 5681 negative sentences extracted from 934 abstracts. 80% of each corpus was used for training with 20% reserved for final blind testing. The training process involved 10-fold cross-validation, i.e. at each iteration, 90% of the training instances were used to build a large number of models and 10% were used to estimate their performance. Extensive sentence classification experiments were performed involving different feature representations, learning algorithms (e.g. neural networks, decision trees, Naïve Bayes classifiers, K-nearest-neighbours), and different SVM kernels and hyperparameter values. Linear SVMs with a C parameter value of 0.1 performed best (precision/recall values of 62/70, 53/66 and 60/69% for structure, function and disease, respectively, when evaluated using the final blind testing sets) and hence were used in METIS.

The second classification component, BioIE (Divoli and Attwood, 2005), uses manually predefined templates and rules to identify sentences relating to the categories of interest. Users may extract

*To whom correspondence should be addressed.

Table 1. Sentence classification results

Topic	Precision (%)		Recall (%)		*n*
	SVMs	BioIE	SVMs	BioIE	
Structure	51 ± 3	33 ± 3	74 ± 3	85 ± 3	20
Function	31 ± 3	16 ± 2	61 ± 5	91 ± 2	18
Disease	48 ± 7	56 ± 5	56 ± 7	79 ± 6	16

all the sentences from each category, or specify keywords to refine the extraction. A link to GPSDB (Pillet *et al*., 2004) allows users to extend their search terms by seeking possible protein synonyms. The templates and user-specified keywords are marked up on the selected sentences, which are in turn ranked according to the number and type/complexity of templates found in them.

3 RESULTS

To further evaluate the performance of the sentence classification components, 20 sets of abstracts were generated by running UniProt sequences through METIS. Precision and recall values were then calculated. An overview of the results is given in Table 1. A list of identifiers used and the full results obtained are available at http://umber.sbs.man.ac.uk/dbbrowser/metis/supp_inf_ids.html and http://umber.sbs.man.ac.uk/dbbrowser/metis/supp_inf_results.html

4 DISCUSSION

Annotating protein sequences and families is onerous and time-consuming, typically involving BLAST searches to gather sequences related to a query, examining hits to primary databases, finding papers and scrutinizing them for relevant information. METIS has been designed to do this automatically. Moreover, it is easy to use since it requires only a single sequence or an ID as input.

By using the biomedical literature cited in Swiss-Prot, METIS circumvents the problems associated with finding relevant publications automatically — effectively, the Swiss-Prot curators have already performed this task for us manually. This approach is similar to that underpinning the MedBlast literature-mining tool (Tu *et al*., 2004); however, MedBlast does not extract any information from the literature it finds. The ability to suggest wider search terms and run the sentence classifiers on any gathered abstracts means that although METIS is Swiss-Prot-based, it is not constrained by the database — should the cited literature prove too narrow in scope or out of date, further information can be gathered and analysed very easily.

Use of online abstracts rather than full texts was a pragmatic choice, as accessing full texts online often involves licensing and subscription issues. Similarly, performing sentence classification rather than information extraction (IE) proper was a practical decision — sentence classification can yield useful results quickly and itself provides an appropriate foundation for true IE, a far more difficult task that is being tackled in BioMinT. Meanwhile, lists of extracted sentences are helpful to annotators, as they are usually concise, semantically complete entities that can be used directly to augment core PRECIS reports.

Our results show that the sentence classifiers embedded in the system perform differently, depending on the sentence types evaluated, e.g. under the test conditions, BioIE performs better at classifying disease-related sentences than the SVM component (precision 56 versus 48%), while for structure-related sentences the opposite is true (precision 33 versus 51%). This finding validates our use of multiple extraction techniques.

The relatively low precision of the function sentence classifiers is disappointing and clearly requires improvement. A likely problem is that the terms used to convey functional information in these sentences are polysemic (not specific to descriptions of function alone) compared with those for structure and disease. Although further SVM training and revision of the function templates may help to improve precision, some syntactic analysis of function sentences will probably also be required to classify them correctly.

Currently, the BioIE component of METIS is flexible—its precision can be increased by manually supplying specific search terms (e.g. a protein name) so that only sentences containing those terms are considered for sentence classification. We are now exploring how we might extend it to perform such specific extraction automatically, using the query terms already suggested by the system for wider literature searching.

METIS is a significant next step towards the automation of database annotation: it reduces the time required to seek out and read relevant literature; it is versatile, yet easy to use; and its output is English-like (being made from Swiss-Prot information and sentences extracted directly from the literature), rendering it immediately useful in the annotation process. As we continue to enhance its performance, the value of METIS will therefore grow as an annotator's assistant.

ACKNOWLEDGEMENTS

This work was supported by European Commission grant number QLRI-CT-2002-02770 BioMinT, EPSRC grant number GR/R80810/01 and the Swiss SER.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Apweiler,R. *et al*. (2004) UniProt: the Universal Protein Knowledgebase. *Nucleic Acids Res*., **32**, D115–D119.

Divoli,A. and Attwood,T.K. (2005) BioIE: extracting informative sentences from the biomedical literature. *Bioinformatics*, **21**, 2138–2139.

Gaizauskas,R. *et al*. (2003) Protein structures and information extraction from biological texts: the PASTA system. *Bioinformatics*, **19**, 135–143.

Mitchell,A.L. *et al*. (2003) PRECIS—an automatic tool for generating protein reports engineered from concise information in Swiss-Prot. *Bioinformatics*, **19**, 1664–1671.

Pillet,V. *et al*. (2005) GPSDB: a new database for synonyms expansion of gene and protein names. *Bioinformatics*, **21**, 1743–1744.

Tanabe,L. *et al*. (1999) MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques*, **27**, 1210–1214 1216–1217.

Tu,Q. *et al*. (2004) MedBlast: searching articles related to a biological sequence. *Bioinformatics*, **20**, 75–77.